

Shadowing the rotating annulus. Part II: Gradient descent in the perfect model scenario

Roland M. B. Young^{a,b,*}, Roman Binter^a, Falk Niehörster^{a,c}, Peter L. Read^b, Leonard A. Smith^{a,d}

^aCentre for the Analysis of Time Series, London School of Economics, London, UK

^bAtmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, UK

^cBermuda Institute of Ocean Sciences, St. George's, Bermuda

^dPembroke College, Oxford, UK

Abstract

Shadowing trajectories are model trajectories consistent with a sequence of observations of a system, given a distribution of observational noise. The existence of such trajectories is a desirable property of any forecast model. Gradient descent of indeterminism is a well-established technique for finding shadowing trajectories in low-dimensional analytical systems. Here we apply it to the thermally-driven rotating annulus, a laboratory experiment intermediate in model complexity and physical idealisation between analytical systems and global, comprehensive atmospheric models. We work in the perfect model scenario using the MORALS model to generate a sequence of noisy observations in a chaotic flow regime. We demonstrate that the gradient descent technique recovers a pseudo-orbit of model states significantly closer to a model trajectory than the initial sequence. Gradient-free descent is used, where the adjoint model is set to $\lambda \mathbf{I}$ in the absence of a full adjoint model. The indeterminism of the pseudo-orbit falls by two orders of magnitude during the descent, but we find that the distance between the pseudo-orbit and the initial, true, model trajectory reaches a minimum and then diverges from truth. We attribute this to the use of the λ -adjoint, which is well suited to noise reduction but not to finely-tuned convergence towards a model trajectory. We find that $\lambda = 0.25$ gives optimal results, and that candidate model trajectories begun from this pseudo-orbit shadow the observations for up to 80 s, about the length of the longest timescale of the system, and similar to expected shadowing times based on the distance between the pseudo-orbit and the truth. There is great potential for using this method with real laboratory data.

This paper was originally prepared for submission in 2011; but, after Part I was not accepted, it was not submitted. It has not been peer-reviewed. We no longer have the time or resources to work on this topic, but would like this record of our work to be available for others to read, cite, and follow up.

Keywords: Shadowing; Rotating annulus; Gradient descent; Numerical Weather Prediction; Data assimilation; Perfect Model Scenario

1. Introduction

Shadowing trajectories are model trajectories consistent with a sequence of observations of a system, given the distribution of observational noise. The existence of such trajectories is a desirable property of any forecast model. If a model does not admit such trajectories then there is no initial condition that remains close to the observations.

The time over which weather and climate models can shadow observations of past weather and climate is unknown. This is concerning given the weight in decision-making that is placed upon output from these models. Techniques for finding shadowing trajectories are, to some extent, understood in low-dimensional systems such as the Lorenz (1963) equations and the Ikeda (1979) map (Judd and Smith, 2001; Judd, 2003; Du, 2009; Smith et al., 2010). There is significant interest in their application to high-dimensional situations such as General Circulation Models (GCMs).

Gradient descent of indeterminism (Judd, 2003; Judd et al., 2008; Stemler and Judd, 2009) is one such technique well-established for finding shadowing trajectories in low-dimensional analytical systems. It starts from a

*Corresponding author. Current address: College of Science, UAE University, P.O. Box 15551, Al Ain, United Arab Emirates. *Email address:* roland.young@uaeu.ac.ae.

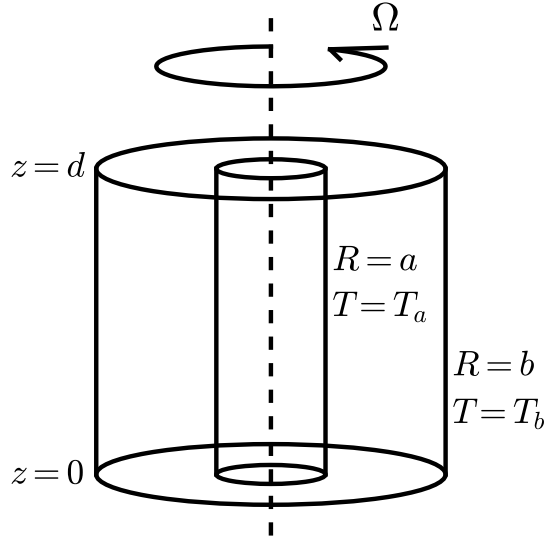


Figure 1: Schematic of a rotating annulus experiment used in the AOPP fluid dynamics laboratory. The inner and outer cylinders at radii $R = a, b$ are at temperatures T_a and T_b respectively. The apparatus rotates at constant angular velocity Ω , and has fluid between the two cylinders.

sequence of observations and alters this sequence by “descending” towards a model trajectory. Alterations to each state in the sequence are calculated based on mismatches (forecast errors) between that state and model forecasts forwards and backwards in time from adjacent states mapped on to the state of interest.

Each state in a sequence constructed by gradient descent is known as a *shadow analysis* (Judd *et al.*, 2008), and in practice this sequence will be a pseudo-orbit of the model (Bowen, 1975) rather than a trajectory. These shadow analyses serve as initial conditions for *candidate* model trajectories. There is no *a priori* guarantee that these candidate trajectories will shadow the original observations, but work with analytical systems has shown gradient descent to be a good method for finding such shadowing trajectories in very low dimensional systems (Smith *et al.*, 2010, for example). The underlying reason for this is not yet well understood. The maximum time that, among all possible candidates, the model shadows the observations for is called the *shadowing time* for that model-observation pair. In practice there are several definitions of shadowing time relevant in different contexts. The ι -shadowing time (Gilmour, 1998, p.47) is the maximum time the distance from model trajectory to observations remains within a bound given by observational error, whereas a ϕ -shadowing time requires the match between model and observations only to be “useful” (Smith, 2000, p.52). In its original sense shadowing refers to the time a true solution of a differential equation remains within a fixed distance of a numerical solution (Bowen, 1975). While superficially similar, the distinction is fundamental and we do not consider this so-called ϵ -shadowing in this work.

Gradient descent itself is a method of noise reduction that has been used in nonlinear and chaotic systems for many years (Kostelich and Yorke, 1988; Grebogi *et al.*, 1990; Hammel, 1990; Farmer and Sidorowich, 1991). It has only recently been applied to higher-dimensional systems, by Judd *et al.* (2004) using an idealised quasi-geostrophic model and by Judd *et al.* (2008) using the US Navy NOGAPS weather model. In neither case were the results used to measure shadowing times against observational data. Its ability to find candidate trajectories that shadow observations for a long time in high-dimensional models of real systems is not yet well explored. Gradient descent has several theoretical advantages over methods in current operational use such as 4D-Var and various flavours of the Kalman filter (Stemler and Judd, 2009; Judd and Stemler, 2010). In low-dimensional test cases it has performed favourably against the extended Kalman filter (Judd, 2003), 4D-Var (Stemler and Judd, 2009), and particle filters (Judd and Stemler, 2009). Its main disadvantage is computational; it is an iterative procedure and $O(100)$ passes through the sequence by the model are required during gradient descent.

Laboratory experiments are intermediate in model complexity and physical idealisation between analytical and global systems. The laboratory setting allows investigation of the properties of gradient descent using a real physical system and a non-idealised model in a situation where the complexity of the flow can be controlled, the experiments can be repeated, and there is potential for long-range observations under laboratory conditions. Whereas an analytical system may have $O(10)$ variables and a general circulation model of the Earth’s atmo-

sphere may have $O(10^7)$, models of laboratory experiments have a more manageable $O(10^4 - 10^5)$ variables. The thermally-driven rotating annulus (Fig. 1) is a classic laboratory experiment representing the mid-latitudes of an idealised generic planetary atmosphere. The “standard” setup uses two cylinders mounted on a turntable, with coincident axes of rotation. Fluid fills the space between the cylinders, which are enclosed by two water baths. Hot water (relative to the working fluid) is circulated around the outside of the outer cylinder as a heat source, and cold water is circulated around the inside of the inner cylinder as a heat sink. The turntable is rotated, usually anticlockwise. This setup mimics the three major influences acting on a planet’s atmosphere: the effects of rotation, gravity, and the temperature difference between low and high latitudes. The annulus exhibits a wide range of dynamical flow regimes describing quasi-periodic, chaotic and turbulent flow, and has become well-established over 50 years as a good laboratory analogue for certain kinds of atmospheric phenomena (*Hide, 1953; Hide and Mason, 1975; Read et al., 1992*).

Since its early development in the 1950s the annulus has been used to conduct research into the fundamental physical processes underlying weather and climate. In recent years effort has also been directed towards using it to inform the development of methods used for weather and climate forecasting. Under laboratory conditions, properties of a particular method can be studied in isolation but using a real fluid as opposed to idealised analytical models more commonly used when testing new methods. There are also several advantages of the laboratory setting compared with atmospheric studies: the controlled nature of the experiment, the degree of reproducibility of the results, and the avoidance of many of the problems associated with atmospheric observations such as a geographically variable observational data density. Effort so far has been directed towards the application of data assimilation techniques such as analysis correction (*Young and Read, 2013*), the ensemble Kalman filter (*Ravela et al., 2010*), and the breeding method for ensemble prediction (*Young and Read, 2016*). With a tangent linear and adjoint model of an annulus model one would also be able to test more recent methods for data assimilation such as 4D-Var (*Rawlins et al., 2007*).

Whether a technique such as gradient descent is feasible for use with high-dimensional GCMs can be informed by its study under the controlled laboratory conditions provided by the annulus experiment. Gradient descent is not yet well-established as a practical method for state estimation in atmospheric systems, but by examining its performance in a real but idealised system a better understanding of whether it could be used operationally will be obtained. A timely comparison to make would be with the results obtained by *Young and Read (2013)* using the well established analysis correction method.

In this paper we demonstrate the gradient descent technique using a model of the rotating annulus under the controlled conditions afforded by the perfect model scenario. We explore how the results depend on the major tuneable parameter in the algorithm, and calculate shadowing times from candidate trajectories produced by gradient descent using the definition presented in *Young et al. (2019, hereafter Part I)*. In the future we intend to extend the work to laboratory data, and compare how long our model shadows reality using gradient descent compared with other assimilation methods. *Gilmour (1998)* attempted to shadow temperature measurements of the annulus using a radial basis function model (but not using gradient descent), but this would be the first attempt to do so using a “full” model of this experiment.

The paper is arranged as follows. In Sect. 2 we describe our simulation of the rotating annulus. In Sect. 3 the gradient descent method is described and its application to the rotating annulus situation is detailed. Section 4 shows the results from our perfect model experiments, and shadowing times are calculated in Sect. 5. The results are discussed and conclusions are drawn in Sect. 6.

2. The rotating annulus model

The mathematical model used to simulate the rotating annulus experiment shown in Fig. 1 is the Met Office / Oxford Rotating Annulus Laboratory Simulation (MORALS) (*Farnell and Plumb, 1976; Hignett et al., 1985; Read et al., 2000*). The model is well established as a quantitatively accurate model of annulus flow in regular and weakly chaotic flow regimes; its details are given in the Appendix. The annulus setup is the “small annulus” configuration used by *Hignett et al. (1985)*. The model configuration is essentially the same as Part I, and the flow simulated is also taken from that paper: rotation rate $\Omega = 1 \text{ rad s}^{-1}$ and temperature difference between the cylinders of $\Delta T = 4 \text{ degC}$. The dimension of the model is $N = 24192$. With this setup and model resolution the general flow behaviour is shown in Fig. 2, and the simulation displays chaotic dynamics. Table A.1 lists the annulus and MORALS parameters.

3. Gradient descent of indeterminism

Consider a sequence of states of a dynamical system x_i valid at times t_i , $i = 0, \dots, w$, where w is the *window width*, and a model f that maps the state x_i forward in time from $t = t_i$ to t_{i+1} . Let each state have dimension N ,

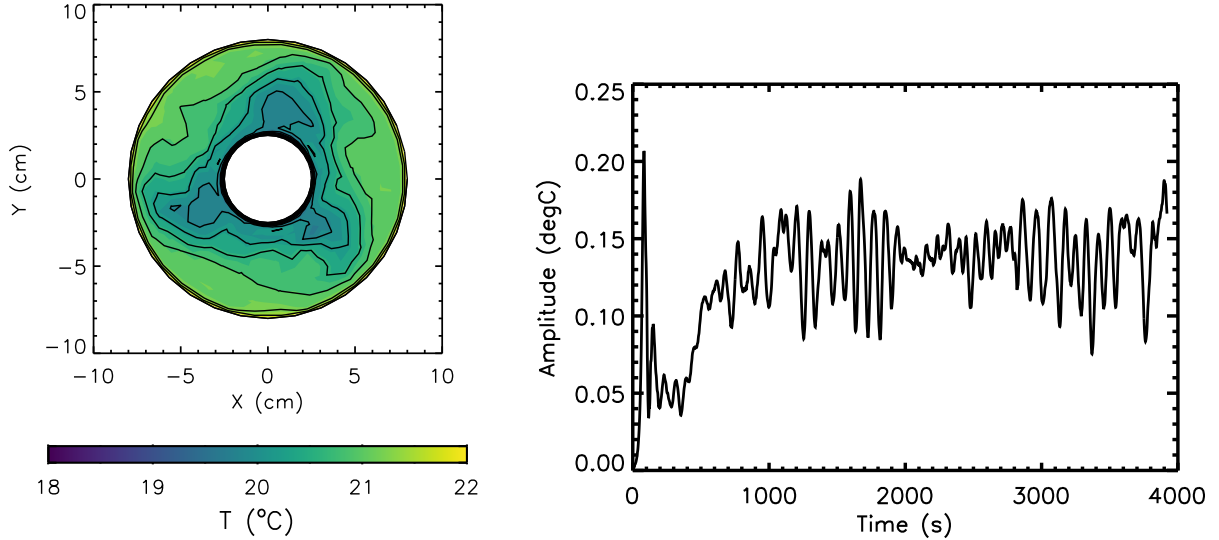


Figure 2: General flow appearance for the setup described in Sect. 2. Top: Horizontal snapshot through the temperature field at $z = 5.61$ cm after 2260 s of simulation (contours every 0.2 °C). Bottom: Time series of the dominant wavenumber-3 mode amplitude for a 3960 s simulation. The dominant mode amplitude is calculated by taking a Fourier transform over an azimuthal ring at mid-radius / mid-height each second during the run. The form of this time series is characteristic of chaotic annulus flow.

hence the complete sequence defines a single point in $\mathbb{R}^{N(w+1)}$. If

$$x_{i+1} = f(x_i) \quad i = 0, \dots, w-1 \quad (1)$$

then the sequence is a *trajectory* of the model f , otherwise it is a δ -*pseudo-orbit* such that $|x_{i+1} - f(x_i)| < \delta \forall i$ (Bowen, 1975). The *mismatch* between consecutive states is

$$\delta x_i = x_{i+1} - f(x_i) \quad (2)$$

The distance between the sequence and a model trajectory can be quantified using a scalar, the *mean squared indeterminism* (or just the *indeterminism*), which is the mean of the squared mismatches over the whole sequence:

$$I = \frac{1}{w} \sum_{i=0}^{w-1} \|\delta x_i\|^2 \equiv \frac{1}{w} \sum_{i=0}^{w-1} \|x_{i+1} - f(x_i)\|^2 \quad (3)$$

where $\|\cdot\|^2$ is the squared Euclidean norm. *Gradient descent of indeterminism* solves the differential equation

$$\frac{d\mathbf{x}}{d\tau} = -\frac{\partial I}{\partial \mathbf{x}} \quad (4)$$

where $\mathbf{x} = (x_0, x_1, \dots, x_w)$, and $\mathbf{x}(\tau = 0) = \mathbf{s}$ is the initial sequence (raw observations or an analysis, perhaps). This equation defines how to change the sequence \mathbf{x} in such a way that the indeterminism falls most quickly, relaxing the sequence of states onto the attractor of the model f . I is a mathematical construct used to guide the gradient descent algorithm and to measure its progress; in general it does not have a physical interpretation. τ is called the *descent time*, after Ridout and Judd (2002), and an intuitive graphical representation of the mechanism is shown in Judd et al. (2004, Fig. 1).

The algorithm itself is an iterative process. Denote state i in the sequence after h iterations by $x_{i,h}$. Each iteration is a two-step process. First, use the model to compute the forecast image $f(x_{i,h})$ and hence the mismatches δx_i for each $i = 0, \dots, w-1$. Second, update the sequence using a discretization of Eq. (4) (Stemler and Judd, 2009, Eq. 3):

$$x_{i,h+1} = x_{i,h} - \frac{2 \Delta \tau}{w} \times \begin{cases} -\mathcal{A}(x_{0,h})\delta x_{0,h} & i = 0 \\ \delta x_{i-1,h} - \mathcal{A}(x_{i,h})\delta x_{i,h} & 1 \leq i \leq w-1 \\ \delta x_{w-1,h} & i = w \end{cases} \quad (5)$$

where $\mathcal{A}(x_i)$ is the adjoint operator $df(x_i)^\top$ of f , and $\Delta\tau$ is a step length in $\mathbb{R}^{N(w+1)}$. One can see from this definition that the change in each state is influenced by information propagated from earlier times via $\delta x_{i-1,h}$ and from later times via $\delta x_{i,h}$ mapped backwards in time by the adjoint operator. At the ends of the sequence information is propagated in one direction only, so the quality of the final sequence is expected to be poorer at the ends (*Ridout and Judd, 2002, Fig. 3*). h is then incremented by one and the procedure is repeated. The indeterminism can only reach zero in the asymptotic limit as $h \rightarrow \infty$, and then only in the perfect model scenario (see below), so in practice the algorithm is stopped when the indeterminism falls below a pre-defined minimum ϵ , or manually after a certain number of iterations.

3.1. Application to the MORALS perfect model scenario

The perfect model scenario (PMS) provides a useful framework for exploring gradient descent and shadowing in complex systems. It allows many aspects of the experiments to be controlled, and provides a ‘‘best case’’ comparison for future results using laboratory data. The PMS simply means that the model f and the system it is modelling, \tilde{f} , are equivalent. In this work we set up the PMS by taking both model and system as MORALS simulations with the same parameters as Part I.

This scenario offers a number of advantages, the most useful of which is that the true state is known exactly and thus explicit comparisons of forecasts with truth can be made. We can define one model run as the true trajectory of the system, and generate artificial observations from that using a known noise model. The PMS offers the greatest amount of control over the experimental configuration, which allows us to study the properties of the system and algorithm in isolation.

In the PMS, gradient descent will converge to a model trajectory under certain conditions (*Ridout and Judd, 2002, Proposition 2*). *Judd et al. (2004)* proved the surprising result that gradient descent still works for incomplete or even wholly unrealistic adjoint operators. In particular, they showed that setting $\mathcal{A} = \lambda \mathbf{I}$, where λ is a scalar and \mathbf{I} is the identity matrix, is sufficient. They call this *gradient-free descent*, as the method can then be used without knowledge of the gradient of the operator f . The conceptual change compared with using the true adjoint is that the algorithm now moves the state in a direction of decreasing indeterminism, but not in the direction of *steepest* descent. We use gradient-free descent here, referring to $\lambda \mathbf{I}$ as the λ -adjoint, because there is currently no adjoint model available for MORALS.

Using this construction we can also test the method with respect to the shadowing properties of candidate trajectories produced from its output. We know *a priori* that the true shadowing time is the whole observation sequence because we already know a model trajectory exists that shadows the observations: the true trajectory the observations were generated from. We expect the candidate shadowing times to be sensitive to how close they begin from that true trajectory, although this does not always hold in practice.

The specifics of how the method was applied to MORALS are detailed in the Appendix, as it differs from the general outline above primarily in matters of notation. In what follows we denote, in general, a state of the annulus model defined on the MORALS grid by the vector \mathbf{x} with $\dim(\mathbf{x}) = N$. We use X to denote, in general, a sequence of such model states:

$$X \equiv (\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_w) \quad (6)$$

for a sequence of length $w + 1$, and hence $\dim(X) = N(w + 1)$, where the states are valid at times t_i . We denote by $\mathbf{x}_{i,h}$ a single state in the sequence (or *shadow analysis*) after h completed iterations of the gradient descent algorithm. X_h is the whole sequence of $w + 1$ states (the *sequence of shadow analyses*) after h gradient descent iterations, and hence X_0 denotes the sequence of observations. We denote the true sequence by \hat{X} , a single state in that true sequence by $\hat{\mathbf{x}}_i$, and a generic true state by $\hat{\mathbf{x}}$.

4. A demonstration of gradient descent using the annulus

We now present a demonstration of gradient descent using the annulus, followed by an analysis of the λ parameter (the parameter that multiplies the identity matrix in gradient-free descent). In the next section we calculate shadowing times for candidate trajectories using the shadow analyses to generate initial conditions. The gradient descent was started from an observational sequence of 65 states separated by 5 s, and ran with $\lambda = 0.5$ for 500 iterations. Other parameters are listed in the Appendix. The sequence is long enough to cover about four periods of the longest timescale associated with the flow. Observations were generated by adding normally-distributed random numbers to a sequence of true states calculated using the model. The random numbers represent observational error of standard deviation $\sigma = 1/3$ of the natural variability of the model at each grid point, denoted by \mathbf{r} . The method used to calculate \mathbf{r} is detailed in the Appendix.

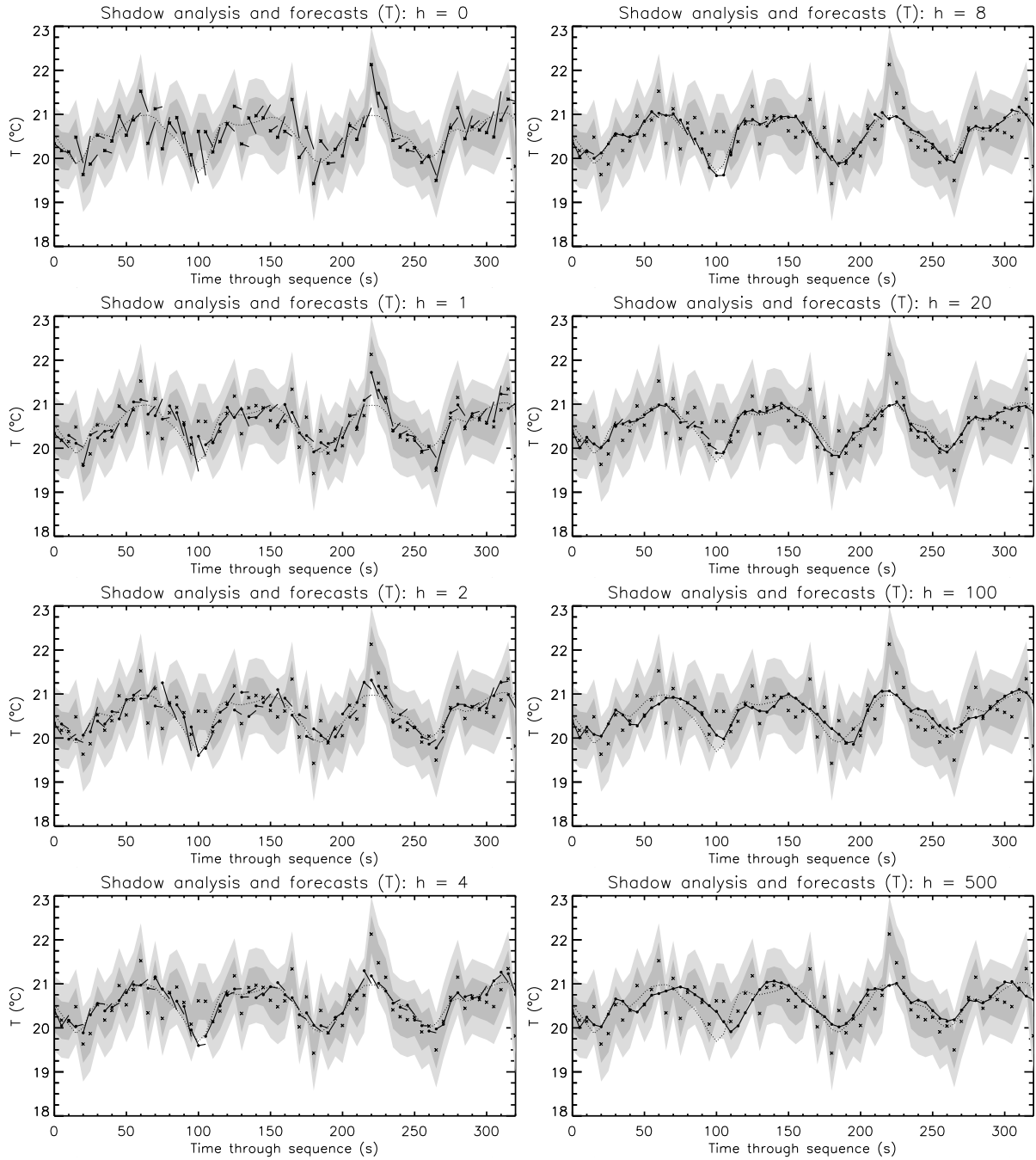


Figure 3: Progression of the gradient descent at a single point. Each panel shows the temperature at a single grid point in the model ($R = 4.81$ cm, $\theta = 6.18$ rad, $z = 5.61$ cm) as a function of time over the gradient descent window. The sequence is shown at the following iterations: $h = 0, 1, 2, 4, 8, 20, 100,$ and 500 . Dots show the shadow analyses, crosses show the observations, solid lines join each shadow analysis at time t_i with its forecast image at time t_{i+1} , grey shaded areas show the range spanned by the observation $\pm 1\sigma$ error (darker shade) and $\pm 2\sigma$ error (lighter shade), and the dotted line is the truth (unknown to the gradient descent algorithm, but included for comparison with the final shadow analysis sequence and the original observations).

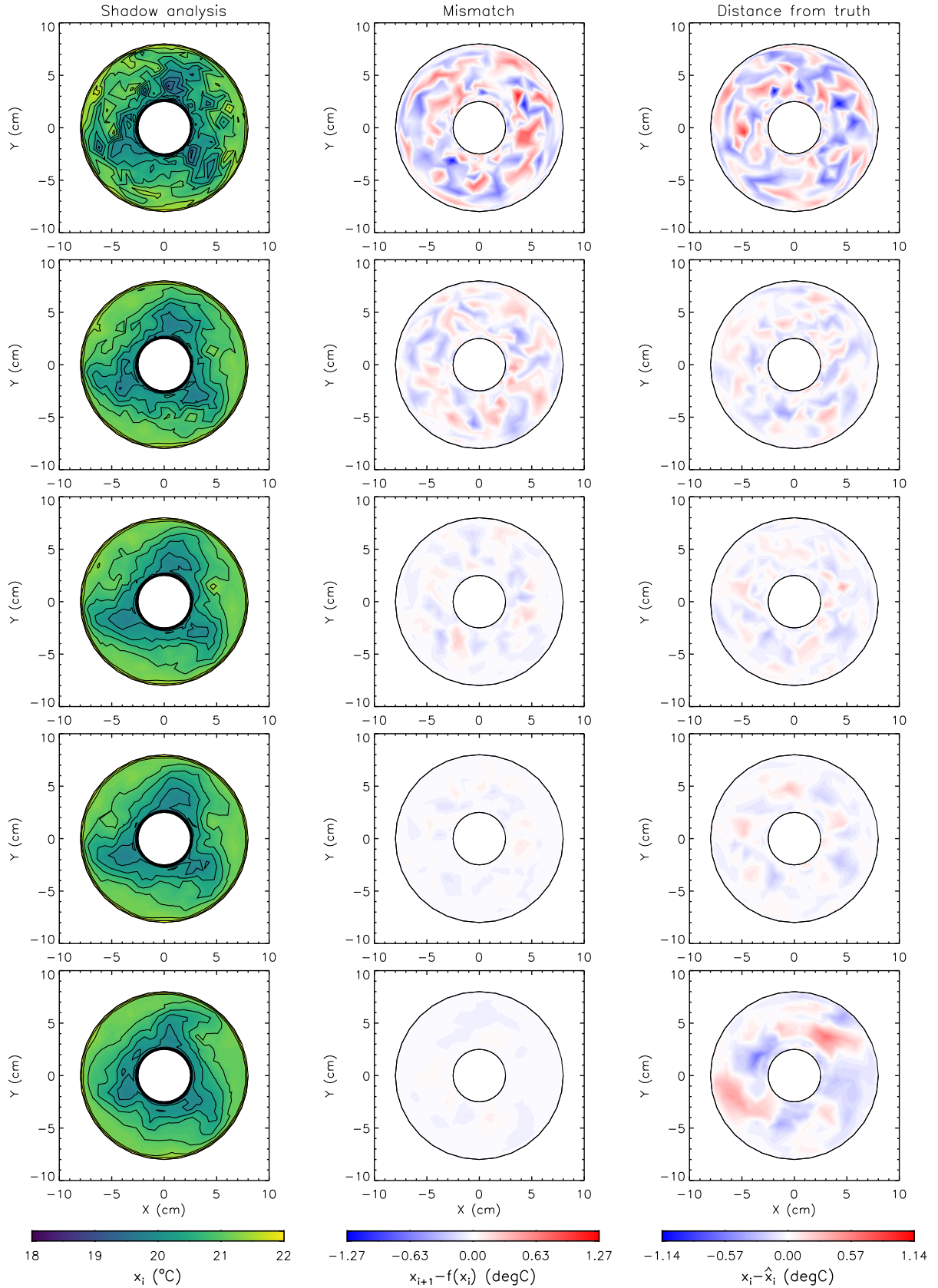


Figure 4: Progression of the gradient descent shown as a horizontal slice through the temperature field at $z = 5.61$ cm and $t = t_0 + 160$ s ($i = 32$, the mid-point in the sequence). The gradient descent is shown at the following iterations: $h = 0, 2, 4, 8$, and 500 (top to bottom). Three quantities are shown: (left) the shadow analysis $\mathbf{x}_{i,h}$ after h gradient descent steps, (middle) the forecast mismatch $\delta\mathbf{x}_{i,h}$ corresponding to that shadow analysis, and (right) the difference between the shadow analysis and the truth $\mathbf{x}_{i,h} - \hat{\mathbf{x}}_i$. The left panel of Fig. 2 shows the truth for this particular run.

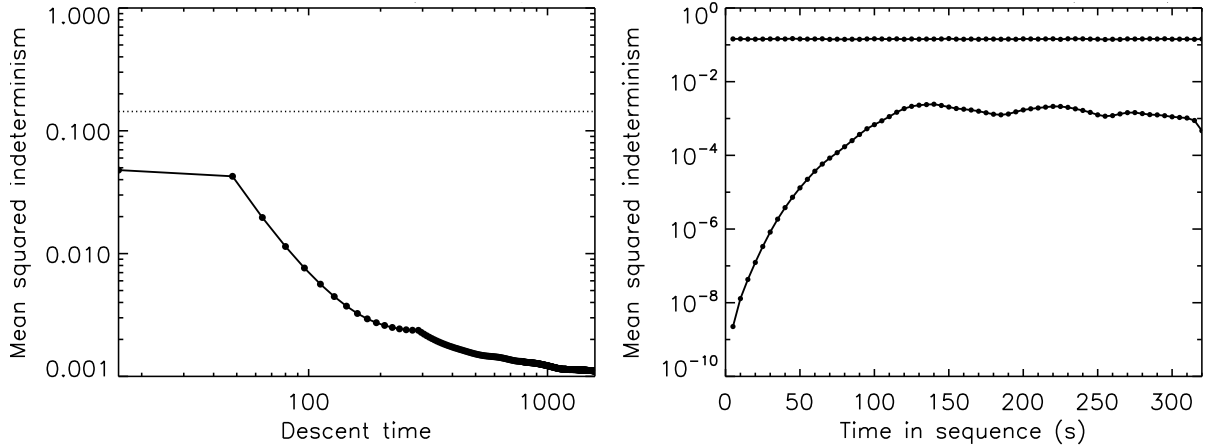


Figure 5: Mean squared indeterminism for the gradient descent in Figs 3–4. The left panel shows the total indeterminism $I(\mathcal{X}_h)$ (Eq. A.6) as a function of descent time τ . Each iteration is represented by a dot, and the horizontal dotted line shows the value at $h = 0$. The right panel shows, as a function of position in \mathcal{X}_h , the mean squared indeterminism for each state $I(\mathbf{x}_{i,h})$ (Eq. A.9) at the start of the gradient descent ($h = 0$, upper line) and at the end of the gradient descent ($h = 500$, lower line).

4.1. Visual demonstration of the gradient descent

In Figs 3 and 4 we show how \mathcal{X}_h changes during the gradient descent. Figure 3 shows the progression of the time series at a single grid point, while Fig. 4 shows a horizontal section through the annulus temperature field near mid-height. These quantities are shown at several steps during the gradient descent. The PMS allows us to see directly how well gradient descent recovers the initial true trajectory $\hat{\mathcal{X}}$, while using information only from the model and observations.

It is easiest to visualise the progress of the gradient descent using the time series in Fig. 3. At the start of the gradient descent ($h = 0$) the states and their forecast images (black dots and solid lines) do not join up at all; \mathcal{X}_0 is far from a model trajectory. During the first several steps of the gradient descent \mathcal{X}_h falls quickly towards a model trajectory. The top panel of Fig. 5 shows $I(\mathcal{X}_h)$ as the gradient descent progresses, as a function of the descent time

$$\tau(h) = \sum_{j=0}^{h-1} \Delta\tau_j \quad (7)$$

Indeterminism falls off approximately as a power law during the first few gradient descent steps. By $h = 8$ the time series in Fig. 3 is close enough to the trajectory to require closer inspection to confirm the time series is not a trajectory, but a pseudo-orbit.

Figure 4 complements the time series in Fig. 3 by showing how the gradient descent progresses over a whole horizontal section. Three different quantities are shown in Fig. 4 as horizontal sections: the shadow analysis $\mathbf{x}_{i,h}$, mismatch $\mathbf{x}_{i+1,h} - f(\mathbf{x}_{i,h})$, and distance from truth $\mathbf{x}_{i,h} - \hat{\mathbf{x}}_i$, all for $i = 32$, the mid-point in \mathcal{X}_h . First, from the middle of these panels we see that the most striking change over the gradient descent is the mismatch. Like the indeterminism in Fig. 5, this falls off very quickly during the gradient descent such that by $h = 8$ the mismatch is some two orders of magnitude smaller than at $h = 0$. This decrease is reflected in the colour scale in that figure. Second, the shadow analysis $\mathbf{x}_{i,h}$ on the left starts off quite noisy and by the bottom of the figure much of the noise has been smoothed out. Finally, on the right the distance between the shadow analysis and the truth also falls rapidly, although not as fast as the indeterminism. Unlike the fall in indeterminism, however, the distance from the truth does not fall off monotonically but begins to increase again by $h = 20$; we shall examine this in more detail below.

Figures 3 and 4 demonstrate the ability of gradient descent to recover a sequence of states much closer to a model trajectory than the original \mathcal{X}_0 in less than ten iterations. While $\hat{\mathcal{X}}$ is included in Fig. 3, the gradient descent algorithm has no information about the truth at all, only the observations and the model.

To check the reproducibility of these results, we ran nine additional gradient descents using exactly the same setup except with different random numbers used to generate the observations. All ten cases produced very similar results, with no outliers. We also ran two additional gradient descents started from different points in the model’s state space, but with an otherwise identical setup. Again, the results were very similar. This reinforces

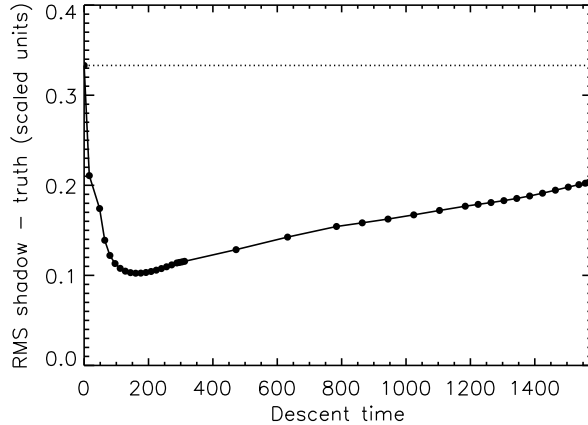


Figure 6: RMS Euclidean distance per grid point scaled by \mathbf{r} between the sequence of shadow analyses and truth (Eq. 8) as a function of the descent time τ . The distance is plotted for each iteration up to $h = 20$ and thereafter for iteration numbers divisible by 20. The horizontal dotted line shows the value at $h = 0$ for ease of comparison with later values.

the conclusion that the initial part of the gradient descent primarily removes noise from the observations, and the latter part converges towards a trajectory; in each case the noise statistics are the same, but the results only diverge once indeterminism has fallen by two orders of magnitude.

Figure 5 shows that, overall, \mathcal{X}_{500} is two orders of magnitude closer to a trajectory than the original \mathcal{X}_0 , both from the full indeterminism in the top panel of that figure, and by comparing the state-wise indeterminism between \mathcal{X}_0 and \mathcal{X}_{500} in the lower panels. At the start of the gradient descent the state indeterminism $I(\mathbf{x}_{i,h})$ (Fig. 5, bottom) is approximately constant with position in the sequence. By construction, the initial expected squared distance from truth at each grid point is approximately $(1/3)^2 \approx 0.11$, the variance of the observational error accounting for scaling by \mathbf{r} .

After 500 gradient descent steps, however, there is a clear structure to the variation of indeterminism with position in the sequence (Fig. 5, bottom, lower line). At the beginning of the sequence there is an approximately exponential growth in indeterminism. This can be explained by the choice of λ . As $\lambda < 1$ more weight is assigned, in the update step in Eq. (5), to $\delta\mathbf{x}_{i-1,h}$ compared with $\delta\mathbf{x}_{i,h}$. Hence more information is passed forwards compared with backwards in time, and so during the gradient descent any given state will tend to reduce the mismatch compared with the state before it faster than the mismatch compared with the state after it. Hence the mismatch at the start of the sequence will decrease the fastest. We shall see later how the value of λ affects this.

4.2. Correspondence between shadow analysis and truth

We have demonstrated that gradient descent recovers a sequence of states close to a model trajectory, but our main reason for its use is that it often produces candidate states with longer shadowing times than other methods used to generate candidates. Above we noted that one trajectory guaranteed to shadow the observations is the original true trajectory, so the correspondence between our pseudo-orbit of shadow analyses and the true trajectory is important. Whether \mathcal{X}_h is close to truth can be measured directly, as we are in the PMS. Figure 6 shows, as a function of descent time τ , the RMS Euclidean distance per grid point between \mathcal{X}_h and the truth $\hat{\mathcal{X}}$ (the distance between two points in $\mathbb{R}^{N(w+1)}$) scaled by \mathbf{r} :

$$D(h) = \left[\frac{1}{N(w+1)} \sum_{i=0}^w \|(\mathbf{x}_{i,h} - \hat{\mathbf{x}}_i) \circ \mathbf{r}^{-1}\|^2 \right]^{1/2} \quad (8)$$

This distance is expressed per grid point so the values are easily comparable with the observational noise standard deviation σ . By construction, the initial distance from truth is approximately $\sigma = 1/3$, as this distance just corresponds to the observational error. The distance from truth then falls off quickly during the first few iterations. After 10–15 iterations, however, the distance from the truth stops falling and begins to rise again, which it does so for the remainder of the 500 iterations. So while $I(\mathcal{X}_h)$ is monotonic during the gradient descent, \mathcal{X}_h approaches $\hat{\mathcal{X}}$ but then moves away again. In this particular example the closest approach to $\hat{\mathcal{X}}$ is about one third of the original distance from it.

This can also be seen in the right panels of Fig. 4. The final three steps show the distance from truth increasing after $h = 8$ (which corresponds to $\tau = 144$, for orientation in Fig. 6), and in the time series in Fig. 3 we can see \mathcal{X}_h

move away from the truth by steps $h = 20, 100, \text{ and } 500$ ($\tau = 312, 864, \text{ and } 1579.1875$). When the mismatch at a single point in X_h is large compared with adjacent points, such as between $t = 70$ and 100 s at $h = 20$ in Fig. 3, this mismatch propagates along the sequence, introducing a phase error in the position of the baroclinic wave along X_h when compared to the values in \hat{X} .

Insight from *Stemler and Judd* (2009) goes some way to explaining this result. The primary reason appears to be the λ -adjoint approximation we are using. They performed a systematic comparison of different adjoint approximations using the *Lorenz* (1963) system, measuring, among other things, indeterminism and distance from truth as a function of computational cost, which corresponds loosely to the number of iterations here. They found that, after about 60 gradient descent iterations using the λ -adjoint (equivalent cost to about 30 steps in our case, as we have 65 states in the sequence while they have 30), the distance from truth began to increase again and eventually diverged (their Fig. 9). They do note that the indeterminism also increases, which is possible in their experiments because they do not change $\Delta\tau$ during the gradient descent, while we do, but they note that the distance from truth begins to increase some time *before* the indeterminism increases.

To explain this result, they note that “the first iterations of the shadowing filter tend to remove effects of observational noise, and subsequent iterations achieve convergence to a trajectory”. In the first iterations a full adjoint doesn’t provide much extra information for the gradient descent, but in the later stages it is vital. *Judd* (2008) examined this more closely using both the PMS and IMS. He argued that (1) when X_h is altered in a way that moves perpendicular to the indeterminism contours the surfaces of constant indeterminism “are well-behaved with smooth slow variations”, while (2) when moving at an angle there is a complex local variation in indeterminism close to $I = 0$. He demonstrates this point using a full adjoint model in the context of the PMS (case 1) and IMS (case 2), but the same principle can be applied in the PMS to a comparison of the full adjoint (case 1) and λ -adjoint (case 2), in which indeterminism is decreased but not in the direction perpendicular to the indeterminism contours. Thus when the indeterminism becomes small compared with its initial value, the full adjoint must be used.

Finally, note that in the PMS local minima can only have $I = 0$ (*Judd and Smith*, 2001, Theorem 2), but only one of these minima will correspond to truth. Intuitively, if the direction taken towards a trajectory is the fastest one possible (i.e. by using the true adjoint) then there is a greater probability that the point on $I = 0$ it approaches will be closer to the truth than a point reached by moving the state around more in I -space. This brings in observational noise as a factor; the smaller the noise the more likely the adjoint is to move the state towards the truth in I -space, however the adjoint is approximated.

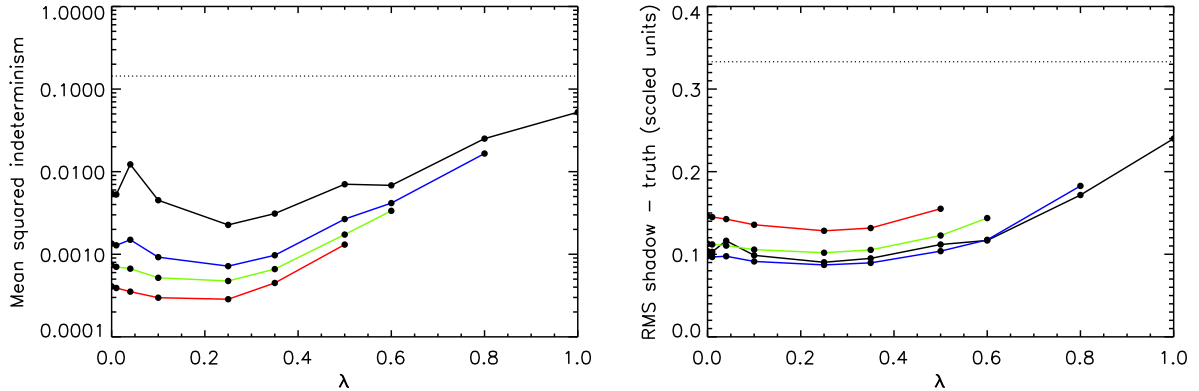
It is encouraging to note that the two orders of magnitude decrease in indeterminism that *Stemler and Judd* (2009, Fig. 9) find for the *Lorenz* system with an analytical adjoint is comparable with the decrease we find in a system of much greater complexity, even though we use the λ -adjoint. Before our results move away from the truth around $h = 10$, this comparison is also true for the distance between shadow analysis and truth. In our case the distance decreases by a factor of three, as does theirs using an analytical adjoint. When they used the λ -adjoint they were only able to decrease the distance to truth by a factor of two. Using the λ -adjoint with a quasigeostrophic model of 1500 variables, *Judd et al.* (2004) were able to produce a descended pseudo-orbit that reduced the initial distance from truth by at most a factor of four (their Fig. 3).

Stemler and Judd (2009) conclude that the optimal strategy is to use the λ -adjoint up to a point while noise is removed, before changing to a more accurate approximation to push the sequence towards a trajectory. With an adjoint model for MORALS this could be done, perhaps changing to the full adjoint once the distance from truth reaches a minimum. This strategy would only be possible in the PMS or IMS, of course; it is not obvious how to determine when the switch should occur when the truth is unknown.

4.3. How the results vary with λ

For a given model setup and observational noise, λ is the primary tunable parameter in the gradient descent. The other parameter, $\Delta\tau$, is optimized by the gradient descent itself. λ quantifies the ratio between mismatch information passed backwards and forwards in time. We ran several gradient descents with exactly the same setup, including the same observations, and ten values of λ : 0, 0.01, 0.04, 0.1, 0.25, 0.35, 0.5, 0.6, 0.8, and 1.

Figure 7 shows how the indeterminism and distance from truth, the two main diagnostics for the progress of the gradient descent, vary with λ during the gradient descent. An intermediate value of λ is optimal; both the rate at which indeterminism falls most quickly and the smallest distance to truth occur at intermediate values. There is a smooth variation in both quantities around the minimum at $\lambda = 0.25$. Increasing λ further degrades both diagnostics: at $\lambda = 1$ indeterminism only falls to half its original value by the end, and the distance from truth is substantially larger than for intermediate λ . In this case the original $\lambda = 0.5$ used by other authors is found to be suboptimal, certainly in terms of indeterminism. The distance from truth reaches a minimum before rising again for all λ . The value of τ corresponding to the minimum distance during the gradient descent varies only weakly as



(a) Indeterminism as a function of λ at the five descent times listed in the caption. (b) Distance between the sequence of shadow analyses and truth.

Figure 7: Progress of the gradient descent (a) indeterminism and (b) distance from truth as a function of λ . Results are shown at four points during the gradient descent: $\tau = 100$ (black), 200 (blue), 400 (green), and 800 (red). The dotted line shows the initial value at $\tau = 0$. Large dots show the values obtained; the lines simply join the dots for clarity. The higher values of λ are not plotted at some of the later descent times because in these cases the step length $\Delta\tau$, which is optimized by the algorithm, had been halved so many times the descent time did not reach the point plotted even after 500 iterations (in some cases $\Delta\tau$ reached double precision rounding error).

λ is varied. In all cases it occurs around $h \approx 10$, between $\tau = 100$ and 200. Furthermore, if $\lambda < 0.5$ then the depth and position of the minimum are only weakly dependent on λ .

In Fig. 8 we show how the distance between the shadow analyses and truth varies along the sequence. Except for $\lambda > 0.6$, the distance from truth over most of the sequence depends very weakly on λ . Each line has a similar structure: at the start of the sequence the distance from truth is largest; it falls exponentially to a constant value; remains approximately constant for most of the sequence, and in some cases falls at the very end. The rate of exponential decay at the start of the sequence is fastest for low λ , and the rate of decrease at the end is fastest for large λ . From a close inspection it was found that, as above, $\lambda = 0.25$ is closest to truth over the part of the sequence where the distance is approximately constant.

The shape of the curves in this plot are very similar to *Judd et al.* (2004, Fig. 3). They saw this effect using a quasigeostrophic model and the λ -adjoint, so this result tells us something more general about the behaviour of the gradient descent using the λ -adjoint. *Ridout and Judd* (2002) argue that at the start of the sequence the distance from truth will decrease exponentially at a rate given by the closest non-positive Lyapunov exponent to zero, and at the end the distance will increase exponentially at a rate given by the smallest non-negative Lyapunov exponent. At the sequence ends information is only passed in one direction, so there is less information there to guide \mathcal{X}_h towards truth. Neither we nor *Judd et al.* (2004) saw the increase at the end of the sequence, however, which they left unexplained. It is not immediately clear why this happens, but perhaps it is because the model itself is not being used to pass information backwards in time. In the λ -adjoint case the propagation of mismatch backwards in time uses no information about the model's Lyapunov exponents, but this information would be included implicitly in the full adjoint.

Later in the gradient descent than Fig. 8 the approximately constant distance from truth over most of the sequence gives way to oscillatory functions of position. In fact the distance between shadow analysis and truth after 500 iterations is quite sensitive to the choice of λ . This change during the later part of the gradient descent implies that during the first iterations the λ -adjoint removes noise rather than searching for the underlying trajectory. Only later does the gradient descent try to converge towards a trajectory, and the underlying variation along the sequence is revealed.

It is not surprising that $\lambda = 0$ does not give the smallest distance from truth. In the trivial case of $\lambda = 0$ no information is passed backwards in time so \mathcal{X}_h will eventually converge to a model trajectory starting from the observed state at the start of the sequence. As the system is chaotic this trajectory will diverge exponentially from truth until it is the same order of magnitude as the model attractor width, so we do not expect $\lambda = 0$ to approach truth no matter how long the gradient descent is run for. Small but nonzero values of λ will also exhibit this effect, but the magnitude of the effect will decrease as λ increases. Hence we expect to find the minimum in the distance from truth at nonzero λ . With indeterminism the effect is similar; at very small λ information is predominantly

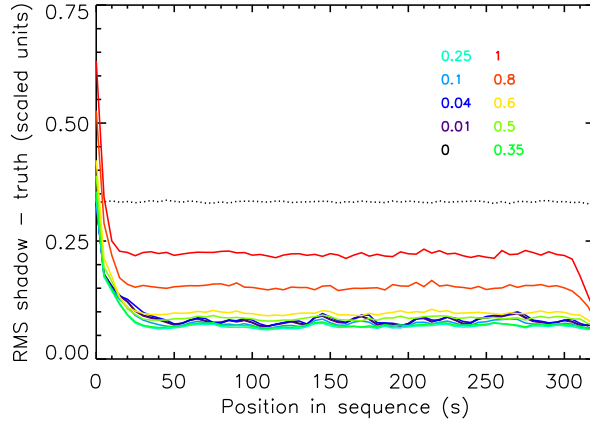


Figure 8: Euclidean distance from each state in the sequence to truth $\|(\mathbf{x}_{i,h} - \hat{\mathbf{x}}_i) \circ \mathbf{r}^{-1}\|$, given as the mean distance per grid point, after $h = 8$ iterations. One line is plotted for each value of λ ; the value for each line is shown in the key. States are separated by 5 s. The dotted line shows the distance from truth at $h = 0$, which is the same for each value of λ as it is a function of observations and truth only.

passed along the sequence in one direction, so only the mismatch between states \mathbf{x}_{i-1} and \mathbf{x}_i pushes the sequence closer to a trajectory. At larger values the sequence is pushed towards a trajectory from both directions, causing the indeterminism to fall more quickly. Like the distance from truth, this effect will increase as λ is increased.

At the other end of the scale, $\lambda \approx 1$, information is passed in both directions equally. Why, then, do the minima not occur at $\lambda = 1$? Here we must consider the effect of approximating the adjoint with a diagonal matrix. If the full adjoint is used the mismatches in both directions contain information about the model at time t_i . When the λ -adjoint is used, however, the information passed backwards in time contains no information about the model at t_i , only at t_{i+1} where the mismatch is calculated. Hence the quality of the update at t_i is suboptimal when the λ -adjoint is used. Hence we might expect the quality of the update to increase as λ is reduced and more weight in the update step is assigned to the model at t_i . Our results show empirically where the balance is between these two effects, at least in the annulus context, and we expect such a trade-off to exist for other chaotic systems, for the same reasons.

The conclusions from this section are clear. First, there is a range of intermediate λ values which give reasonable results both in terms of indeterminism and distance from truth, while at both extremes of the range the quality of the gradient descent is compromised. Second, the distance from truth as a function of position in the sequence confirms something more general about using the λ -adjoint, given the comparison of our results with *Judd et al. (2004)*. There is a trade-off between two mechanisms that degrade the quality of the gradient descent for extreme values of λ . For $\lambda > 0.5$ the quality of the update step is degraded by the relatively large weight assigned to information passed backwards in time sub-optimally. $\Delta\tau$ also falls very quickly with high λ , so only a small amount of descent time is covered, limiting the potential of the gradient descent to proceed much further. For $\lambda < 0.1$ the sequence converges to a trajectory starting from very close to the first observation, so chaos causes the rest of the shadow analyses to diverge from truth. For intermediate values of λ , where the combined effect is minimized, sequences are recovered closest to a trajectory and to truth. We recommend a value around 0.25 for future applications using the λ -adjoint, although within this intermediate range the results are only weakly dependent on λ .

5. Shadowing times from the sequence of shadow analyses

Gradient descent produces, from a sequence of observations, a pseudo-orbit of the model closer to a trajectory than the original observation sequence. Each state on the pseudo-orbit is the start of a candidate trajectory of the model. The shadowing time is the maximum time any of these candidates shadow the observations. We can also generate additional candidates using linear combinations of states on the pseudo-orbit and forecast images of earlier states.

In this section we measure how long the model can shadow observations using candidates from a pseudo-orbit produced by gradient descent in the previous section, each candidate beginning a model trajectory. The *shadowing time* τ_S , is the maximum among all *candidate shadowing times* τ_S starting from candidates \mathbf{x} , i.e. $\tau_S = \max_{\mathbf{x}} \tau_S(\mathbf{x}, t)$. From the previous section we choose candidates from the gradient descent with $\lambda = 0.25$ at

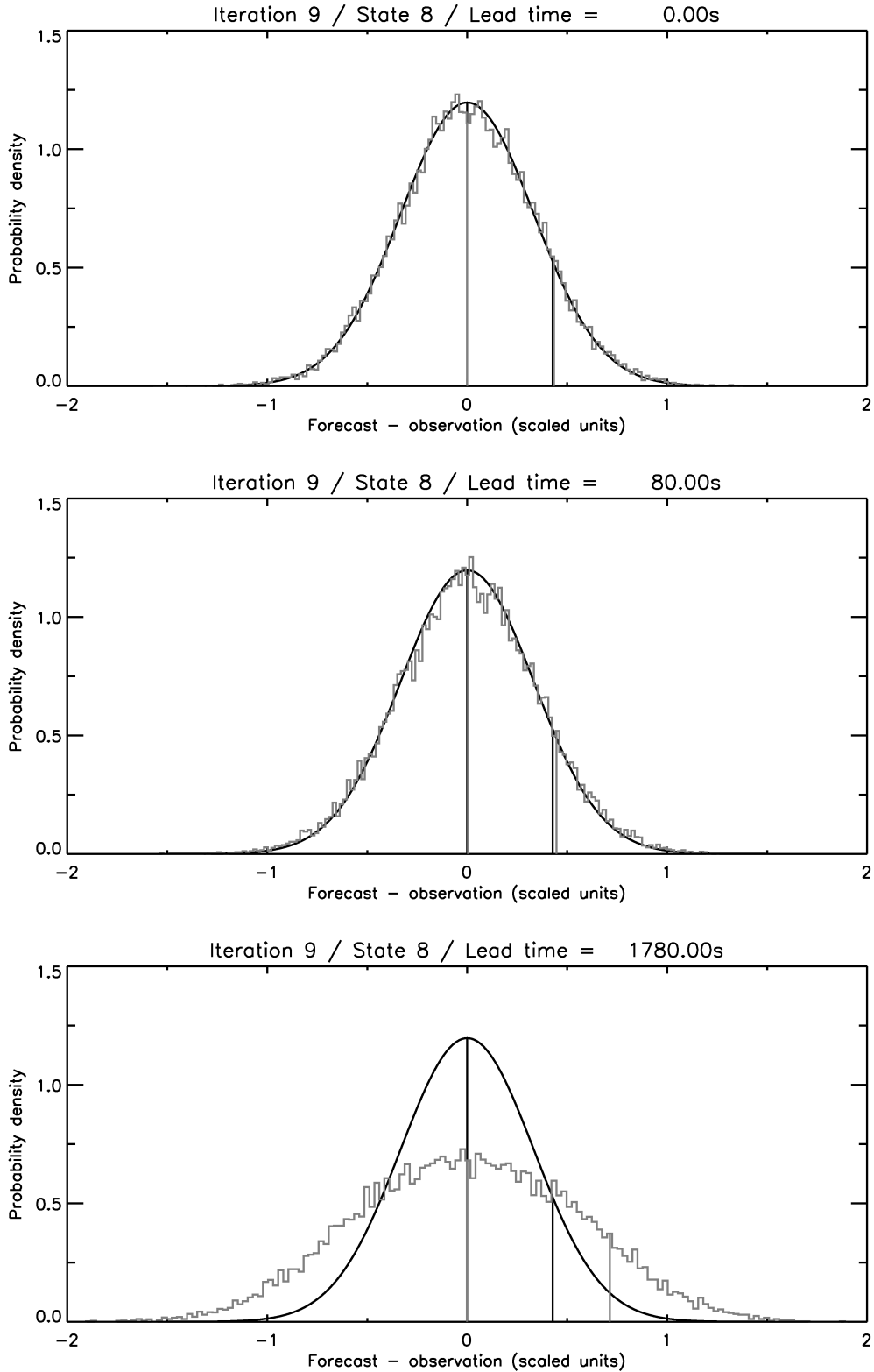


Figure 9: Distributions of residual errors $\mathbf{e}[t]$ (Eq. 9) at three lead times from the candidate trajectory started from state $i = 8$ at $h = 9$ for the gradient descent with $\lambda = 0.25$. Top: $t = 0$, middle: $t = 80$ s (the first time the trajectory fails to shadow observations at significance level $p = 10^{-5}$), and bottom: $t = 1780$ s (the end of the candidate trajectory). The black and grey curves are the noise and residual distributions, and the vertical black and grey lines are the 50th and 90th percentiles of the noise and residual distributions. Notice how strict this definition is — even in the middle panel the residual distribution is only marginally different from the noise distribution, yet the model does not shadow at this time.

iteration $h = 9$, because at that iteration that gradient descent came closest to truth among all the cases that were run. Part I describes the method used to compute the shadowing time; we summarise it below.

Our shadowing time quantifies how well the gradient descent performs against a benchmark set in Part I. In that work they used the same model parameters as in this paper, except here we have used $\sigma = 1/3$ instead of $\sigma = 0.1$. They generated a cloud of candidate initial conditions a fixed distance from the true state and for each candidate measured how long it shadowed a subsequent set of observations. These times provide a benchmark against which we can compare the shadowing times from our gradient descent experiments — the shadowing times we might expect depend on the distance of the candidates from the original true trajectory. The largest initial distance from truth that Part I used was about 17.5% of the observational error, and in that case they measured shadowing times of 50–150 s. In the previous section we found the distance from truth of our best-case result is around 25% of the observational error (Fig. 7), so we might expect a maximum shadowing time around 100 s.

From the selected pseudo-orbit we took as candidates at each t_i (1) the state on the pseudo-orbit $\mathbf{x}_{i,h}$, (2) the state halfway between this state and the forecast image of the previous state, $\frac{1}{2}[\mathbf{x}_{i,h} + f(\mathbf{x}_{i-1,h})]$, and (3) the images of all the previous states on the pseudo-orbit mapped to t_i , for both $\mathbf{x}_{i,h}$ and the halfway state. The shadowing times for the candidates (3) are available at no extra computational cost. Each candidate was used to start a single MORALS trajectory.

5.1. Measuring the shadowing time

Following *Smith et al.* (2010), our candidate shadowing time τ_S for a particular candidate trajectory is the trajectory length over which the residual error distribution remains consistent with the observational error distribution. The vector of residual errors is

$$\mathbf{e}[t] = \{f^t(\mathbf{x}[0]) - \mathbf{s}[t]\} \circ \mathbf{r}^{-1} \quad (9)$$

where $\mathbf{x}[0]$ is the initial candidate state, $\mathbf{s}[t]$ are the observations at time t from the beginning of the candidate trajectory, and f^t denotes integration of the model for time t . We scale the raw residual error by \mathbf{r} so we can combine different physical quantities into one distribution. We test the null hypothesis that the vector $\mathbf{e}[t]$ is a sample drawn from the observational error (noise) distribution $N(0, \sigma^2)$ (the comparable distribution once raw values are scaled by \mathbf{r}). Part I showed that, using order statistics, the distribution of a specific percentile of the noise distribution can be found analytically. We find (empirically) the 50th and 90th percentiles of the residual distribution $\mathbf{e}[t]$, and test whether, at a particular significance level p , these are drawn from the equivalent noise distribution (found analytically). For the candidate trajectory to shadow the observations at time t we require *both* percentiles to fall within the respective confidence intervals of the noise distribution.

If the model shadows the observations at time t by this definition, we proceed to the next observations (5 s later in this case) and repeat the procedure. The shadowing time τ_S for a particular candidate is the last time at which the residual error distribution is consistent with the noise distribution. Figure 9 shows this definition in use.

We require a suitable significance level p to find τ_S . Because our shadowing time algorithm requires multiple significance tests, the probability of a Type I error (rejecting the null hypothesis when it is true, and hence setting an erroneously low τ_S) increases as more tests are done. Part I calculated the maximum p leading to fewer than one expected Type I error over all candidates. The calculation is outlined in Appendix A.5; we found $p = 10^{-5}$ to be sufficient.

5.2. Measured shadowing times

We measured shadowing times τ_S for each of the candidates. Figure 10 shows an example time series from one of the candidate trajectories, and Fig. 11 shows all the candidate shadowing times τ_S . From this plot we can read off $\tau_{S_i} = 80$ s as the shadowing time for this pseudo-orbit of descended states. This is about one quarter of the total length of the original observational sequence, or about one period of the longest timescale of the system, the oscillation of the main baroclinic wave. This time is near the low end of the range of shadowing times obtained in Part I for their largest perturbation from truth, but the descended pseudo-orbit’s distance from truth in this case was slightly larger, around 25% of the observational error compared with 17.5%.

There are no particular trends relating individual candidates’ shadowing times to their positions in the sequence, except the candidates starting from very close to the start of the sequence do not shadow at all. This is not surprising since the quality of the states near the start of the sequence are poorer than in the middle because information is only passed in one direction at the start of the sequence. A few observational states appear to be difficult for the candidate trajectories to shadow (e.g. around 140–145 s), causing all candidate trajectories approaching it to fail to shadow at that point. Within 45 s of the end of the sequence we found candidates that shadow to the end of the sequence.

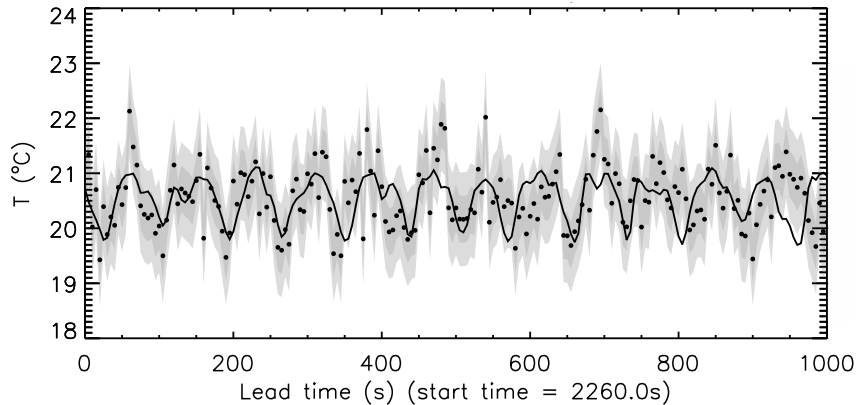


Figure 10: Temperature time series at the single grid point used earlier in the paper (Fig. 3) for the candidate trajectory started from state $i = 32$ in the sequence at $h = 9$ with $\lambda = 0.25$. The solid black line shows the forecast, black dots show the observations, and the shaded area represents observations $\pm 1\sigma$ (darker shading) and $\pm 2\sigma$ (lighter shading).

Our gradient descent setup and algorithm is really just a method for selecting candidate trajectories we think will shadow for a long time. It is always possible that we have missed candidates that shadow for longer, so our measured shadowing time τ_S , can only be a lower bound. Indeed, we know (by construction) that at least one candidate exists that shadows the whole observational sequence: the true sequence that generated the observations in the first place. Clearly our τ_S , does not approach this, being only about a quarter of the original sequence length. However, our shadowing time is consistent with those for candidate trajectories starting from a similar distance from the truth in Part I. In this respect our results are encouraging, because they show that the distance from the candidate state to truth is the primary predictor of the subsequent candidate shadowing time. It shows that better shadowing times will come from improvements to the gradient descent method that bring the pseudo-orbit closer to truth. We identified some possible improvements in the previous section, foremost among these being to use a full adjoint model instead of the naïve λ -adjoint. With a full adjoint we expect to shadow for considerably longer.

6. Discussion and conclusions

We have implemented gradient descent of indeterminism for the thermally-driven rotating annulus in the perfect model scenario. Our results show that a sequence of states much closer to a true system trajectory can be recovered using gradient descent. Diagnostics based on indeterminism and distance from the truth showed our demonstration gradient descent recovered a sequence of states in which the indeterminism had fallen by two orders of magnitude. The sequence converged towards truth as the gradient descent progressed but then moved away from truth once the distance had fallen by a factor of three. This was attributed to the λ -adjoint approximation. An analysis of varying λ showed that the gradient descent is optimized around $\lambda = 0.25$. In that case indeterminism falls by three orders of magnitude after 500 gradient descent steps and the distance from truth falls by a factor of four except near the start of the sequence.

Candidate trajectories started from one particular \mathcal{X}_h were used to obtain shadowing times using the method developed in Part I. We found the model shadows the observations for $\tau_S = 80$ at the $p = 10^{-5}$ significance level. This was at the lower end of the range of times obtained in Part I for candidate trajectories started a similar distance from truth. Our shadowing time is encouraging because the initial distance between \mathcal{X}_h and truth could be decreased further by using a more accurate adjoint model for the gradient descent.

We discussed above how the λ -adjoint causes the sequence of shadow analyses to move away from truth after the first several steps of the gradient descent. Despite this result, we are most encouraged that, even with the λ -adjoint, the distance from truth can be reduced by a factor of four over most of the sequence. The most important next step in this work is to include a full adjoint model for MORALS. At the time the algorithm was implemented no adjoint model was available, but one now exists (Hussain, 2010), which should be a major step forward in using this model for various purposes. This improvement may be of greatest benefit in the PMS, as in experiments with observational data Judd *et al.* (2008) demonstrated that the λ -adjoint may be sufficient because the advantages of a full adjoint in the later part of the gradient descent are offset by the model being an imperfect representation of the system. The attractors of the model and system will be disjoint, and so there may be little to gain from a full adjoint when its main advantage is to navigate through the complex structure of the indeterminism contours

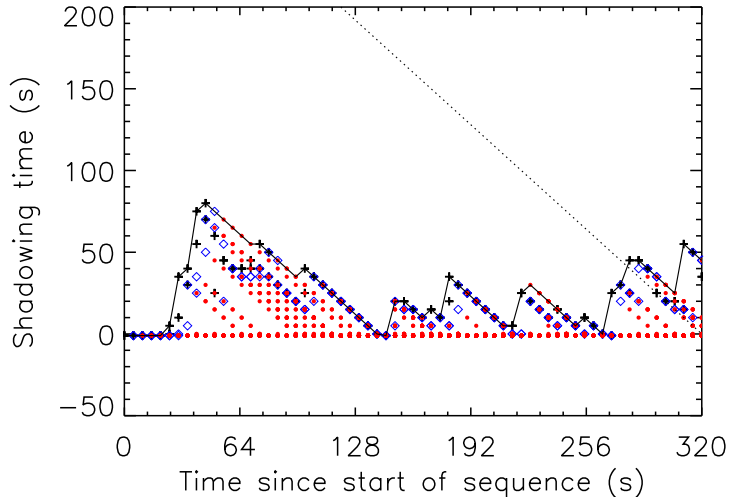


Figure 11: Shadowing times τ_S for each of the candidates described in the text, for the pseudo-orbit after $h = 9$ in the gradient descent with $\lambda = 0.25$. At each position in the sequence, crosses (black) are candidates $\mathbf{x}_{i,h}$ and $\frac{1}{2}[\mathbf{x}_{i,h} + f(\mathbf{x}_{i-1,h})]$, diamonds (blue) are candidates from one-step forecast images of the previous step, i.e. $f(\mathbf{x}_{i-1,h})$ and $f(\frac{1}{2}[\mathbf{x}_{i-1,h} + f(\mathbf{x}_{i-2,h})])$, and dots (red) are candidates from forecast images of all previous states. The solid line traces the maximum τ_S over all candidates at each point in the sequence, and the dotted line separates candidates that shadow to the end of the original sequence of observations (above the line) from those that don't (below it).

near the model attractor. Nevertheless, an accurate adjoint with an imperfect model beats an inaccurate adjoint with an imperfect model. Any increased accuracy gained with the full adjoint using real observations should be measured and compared with the λ -adjoint. Quantification of this increased accuracy would be useful for informing operational implementation of any algorithm based on gradient descent methods. If both the λ -adjoint and full adjoint are used, it is not clear when the switch over should occur when the true state is not available. One option is to switch over when the distance between X_h and X_0 (the *implied noise*) first reaches a maximum.

We are excited by the possibility of applying this method to real annulus laboratory data. Gradient descent allows model error to be examined in a systematic way, by examining the geometric relationship between observations and the model attractor (Judd *et al.*, 2008). Application of this method to laboratory data will allow model error to be explored in more detail, leading to a more fundamental understanding of the limitations of our annulus model. When using observational data the visual progression of the gradient descent (i.e. Fig. 4) is even more revealing, as then one sees how the model adjusts itself where the original sequence of observations or analyses is far from the model manifold. This reveals locations and features in the flow that are poorly simulated by the model, and also how the model attempts to adjust itself to fit to those observations. One particular annulus dataset that might be of particular interest is a wavenumber-3 structural vacillation flow where the observed wave is phase-locked to the tank because of the deposition of tracer particles onto the bottom. This flow is generally poorly modelled by MORALS (Young and Read, 2013), and using this dataset with gradient descent should provide some insight into how the model fails in this case.

The broader aim is to determine whether gradient descent would be feasible in an operational context. For this to be the case, it would (at least) need to out-perform the current state-of-the-art data assimilation technique used in forecasting centres worldwide, 4D-Var (Rawlins *et al.*, 2007), and out-perform the sequential Bayesian methods in development such as the ensemble Kalman filter (EnKF) (Evensen, 1994) and particle filters (van Leeuwen, 2010). While a numerical comparison is beyond the scope of this paper, there are a number of conceptual advantages of gradient descent over these other methods. These differences are discussed in more detail by Stemler and Judd (2009) and Judd and Stemler (2010).

Sequential Bayesian methods like the Kalman filter and its variants suffer from the major problem that the observational noise must be small compared with the system's nonlinearities. Sequential methods also cannot correct poor state estimates in the past whose error then propagates forward (Judd and Stemler, 2010). For nonlinear systems in particular the forward propagation of errors can introduce large errors very quickly. Gradient descent avoids these problems with sequential filtering as it makes no assumptions about the linearity of the system. Indeed, nonlinearity is actively exploited. In addition, it uses information from the past simultaneously with information from the present in each state estimate. Judd (2003) showed that gradient descent compares favourably with the extended Kalman filter (EKF), except when the dynamical noise exceeds observational noise.

An additional problem with these and any methods based on maximum likelihood estimates is that even in the PMS these methods fail to assign that maximum likelihood to the true state of the system (Judd, 2007). Comparing gradient descent with particle filtering using the Ikeda (1979) system, Judd and Stemler (2009) found that gradient descent recovers state estimates closer to truth than the particle filter in almost all cases, even when using an optimized particle filter and “out-of-the-box” gradient descent.

Of the variational methods 4D-Var is, in a sense, also a method that searches for shadowing trajectories. The fundamental difference is that 4D-Var uses a kind of “shooting” method; it alters the initial state in a sequence and compares the model trajectory generated from that initial state over a window of observations. Instead of indeterminism the cost function is (Stemler and Judd, 2009, Eq. 8)

$$C(x) = \frac{1}{n} \sum_{i=1}^n \|s_i - f^i(x)\|^2 \quad (10)$$

The problem with this method is that sensitivity to initial conditions means that the window length over which 4D-Var can be realistically applied is severely restricted. Gradient descent suffers from no such problem as states all the way along the window are used simultaneously, and hence each forecast needs to be optimized only over the time between it and the next state in the sequence. Stemler and Judd (2009, Fig. 8) demonstrate this problem with 4D-Var using the Lorenz (1963) system. A related variational method, weakly constrained 4D variational assimilation (WC4DVA), also has some similarities to gradient descent. Stemler and Judd (2009, p. 1268), Judd (2008, p. 221), and Judd and Stemler (2010, pp. 268–9) argue strongly, however, that these similarities are superficial. In particular, they show that there is an inconsistency between what WC4DVA claims to solve and what the method actually solves.

Gradient descent offers a number of additional practical advantages over these other methods. First, the algorithm is, in our opinion, conceptually simpler than variational or EnKF methods. Second, when using real data the number of tunable parameters is generally less than other assimilation methods. Analysis correction (Lorenz *et al.*, 1991) has about ten, for example. With a full adjoint the only tunable parameter in gradient descent is $\Delta\tau$, and even then its value can be optimized by the gradient descent as described in Appendix A.4. Third, the background error covariance matrix is not required, the calculation of which is generally a major challenge for other methods. Model errors are “discovered, not prescribed” (Judd *et al.*, 2008; Judd and Stemler, 2010), providing information about where and how the model fails to simulate reality.

The main problem we found using gradient descent in this system was the burden on computational resources. Each gradient descent step requires the resources for one complete pass through \mathcal{X}_h with both the forward model and the adjoint model. In the PMS this is a major problem as many hundreds of iterations are required. This would not be such a problem with laboratory data, however, because models of real systems are imperfect. The gradient descent is therefore not expected to converge to a trajectory of the model. Instead it will converge to a pseudo-orbit and in practice the gradient descent is terminated when the standard deviation of the forecast mismatches approaches observational error. This only takes some tens of steps, which is a great improvement over the hundreds of steps required in the PMS. Even so, tens of iterations of a GCM over a lengthy sequence of observations is somewhat more computation than is currently used in operational assimilation. For example, the Met Office 4D-Var scheme uses one pass of the nonlinear model and 5–6 passes of the linearized model through the sequence (Rawlins *et al.*, 2007). For gradient descent to be a feasible operational method, therefore, it would need to be shown that its additional accuracy is worth the computational expense.

With these comparisons in mind, we feel there is great potential for using gradient descent for state estimation in high-dimensional models, and in particular the rotating annulus’ part in developing and testing the method. The framework for using gradient descent in the annulus context is now in place. In the future one could extend it to estimates of shadowing times in the various annulus flow regimes, experiments in the imperfect model scenario, and experiments starting the gradient descent from laboratory data. Several questions present themselves: How long can the laboratory annulus be shadowed in different flow regimes? How long a window is required to shadow with the same accuracy as sequential methods (in particular analysis correction; Young and Read (2013) have produced a set of rotating annulus assimilation results for comparison)? How quickly does the gradient descent converge, and is it quicker when an analysis correction analysis is used to begin the gradient descent? How do these results depend on model resolution, and is there a resolution beyond which no further improvements can be made?

Finally, we intend to use gradient descent as part of a larger programme of research using the annulus as a test bed for meteorological methods in current use and development. As mentioned above, analysis correction has already been implemented by Young and Read (2013) as an example of a well-established assimilation method. We intend to implement and compare some of the other methods discussed in this work in the annulus context, for example 4D-Var and the particle filter; Ravela *et al.* (2010) have already made some progress with the EnKF.

Table A.1: Annulus and MORALS parameters. Fluid properties are parameterized as a function of temperature using the expressions in *Hignett et al.* (1985, Table 1).

Inner cylinder radius	a	2.5 cm
Outer cylinder radius	b	8.0 cm
Annulus depth	d	14.0 cm
Rotation rate	Ω	1.00 rad s ⁻¹
Reference temperature	T_R	22 °C
Inner cylinder temperature	T_a	18 °C
Outer cylinder temperature	T_b	22 °C
Temperature difference	ΔT	4 degC
Fluid		17% glycerol / 83% water by volume
Density	ρ_0	1.043 g cm ⁻³ at 22 °C
Viscosity	ν_0	0.0162 cm ² s ⁻¹ at 22 °C
Thermal diffusivity	κ_0	0.00129 cm ² s ⁻¹ at 22 °C
Model timestep	δt	0.02 s
Radial grid points	N_R	16
Azimuthal grid points	N_ϕ	32
Vertical grid points	N_z	16

Shadowing methods also facilitate ensemble generation using the theory of indistinguishable states (*Judd and Smith*, 2001, 2004), and we believe the application of this particular method in a real physical system would also be a timely comparison to make with current methods for ensemble generation.

Acknowledgments

We thank Daniel Bruynooghe, Hailang Du, Kevin Judd, and Thomas Stemler for useful conversations on a number of topics. RMBY and FN acknowledge financial support from the Grantham Research Institute on Climate Change and the Environment. RMBY also acknowledges financial support from NERC Studentship NER/S/A/2005/13667.

Appendix A. Technical details

Appendix A.1. MORALS

MORALS solves the Navier-Stokes, heat transfer, and continuity equations subject to the Boussinesq approximation, in cylindrical polar coordinates. Four prognostic variables are defined: three velocity directions \mathbf{u} (radial), \mathbf{v} (azimuthal), \mathbf{w} (vertical), and temperature \mathbf{T} . A fifth field required in the prognostic equations is kinetic pressure $\mathbf{\Pi} \equiv \mathbf{p}/\rho_0$, which is diagnostic and is calculated from the other four fields using a Poisson equation. ρ_0 is the fluid density at a reference temperature. The fluid rotates at constant angular velocity Ω , all velocities are set to zero at the boundaries, the temperature gradient is zero across the top and bottom boundaries, and the temperatures at $R = a$ and $R = b$ are T_a and T_b respectively. \mathbf{T} is defined relative to a reference temperature T_R (22 °C here) and $\mathbf{\Pi}$ is relative to a reference pressure $\Pi_0(R, z) = \frac{1}{2}\Omega R^2 + g(d - z)$. The fields are discretized on a staggered Arakawa C grid (*Arakawa and Lamb*, 1977), and are non-uniform in the radial and vertical directions to resolve the boundary layers.

With four prognostic variables using the resolution in the table there are $N_{\text{tot}} = 4N_R N_\theta N_z = 32768$ variables in total. The number of independent variables, N , is less than this because points on and outside the fluid boundary (outside points are required by some boundary conditions) are fixed by values at other grid points. We use N as the dimension of the model, $N = 24192$. When working in the PMS and correspondence with the laboratory experiment is less important, the choice of resolution depends on a balance between the available computer resources and the number of simulations required for that experiment. Because of the large computational overhead required by gradient descent, we have used a lower resolution than is normally used for annulus simulations that are compared with laboratory observations.

Appendix A.2. Experimental parameters

Table A.1 lists the annulus and MORALS parameters. Table A.2 lists the experimental parameters for the demonstration gradient descent in Sect. 4.

Table A.2: Experimental parameters for the demonstration gradient descent in Sect. 4.

Demonstration gradient descent - truth and observations		
Spin-up time	t_{spinup}	2000 s
Pre-sequence time	t_{preseq}	100 s
Time between states	Δt	5 s
Window width	w	64
Time of first state ($i = 0$)	t_0	2100 s
Time of final state ($i = 64$)	t_w	2420 s
Sequence length	$t_w - t_0$	320 s
Observational noise	σ	1/3
Demonstration gradient descent - gradient descent parameters		
Initial step length	$\Delta\tau(\tau = 0)$	16.0
Cut-off indeterminism	ϵ	10^{-28} (machine precision)
Gradient-free descent parameter	λ	0.5
Maximum number of iterations	h_{max}	500

Appendix A.3. *Generating the initial sequence of observations*

A full MORALS state \mathbf{x} is a concatenation of the four fields u , v , w , and T :

$$\mathbf{x} \equiv \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \\ \mathbf{w} \\ \mathbf{T} \end{pmatrix} \quad (\text{A.1})$$

where $\dim(\mathbf{x}) = N = N_u + N_v + N_w + N_T$.

The sequence of artificial observations \mathcal{X}_0 was generated by adding noise $\tilde{\mathbf{e}}_i$ to the true states $\hat{\mathbf{x}}_i$ generated by MORALS. The method is very similar to the method used in Part I. The true sequence is generated by MORALS in three stages: (1) from $t = 0$ to t_{spinup} , (2) from $t = t_{\text{spinup}}$ to $t_0 = t_{\text{spinup}} + t_{\text{preseq}}$, and (3) from $t = t_0$ to t_w . The third stage corresponds to $\hat{\mathcal{X}}$.

The first stage (the ‘‘spin-up’’ phase) is required to spin up the model from rest (in the rotating frame of reference) to a state in which transient behaviour has decayed and a coherent flow structure is present.

The third stage (the ‘‘sequence’’ phase) contains the sequence of states $\hat{\mathcal{X}}$ used to obtain observations to start the gradient descent. The states are separated in time by Δt .

The second stage (the ‘‘pre-sequence’’ phase) is used purely for the generation of observations in the PMS. We use the sequence of states between $t = t_{\text{spinup}}$ and $t = t_w$ to obtain an estimate, at each model grid point, of the range of values admitted by the model when it is in dynamical equilibrium. This range can be interpreted as a measure of the natural variability of the system at each point in space. We denote this range by the vector \mathbf{r} , where each element represents the range of values that encloses 99% of the values over the pre-sequence and sequence phases at that grid point:

$$\mathbf{r} = \tilde{\mathbf{x}}[t_{\text{spinup}}, \dots, t_w]_{99.5\%} - \tilde{\mathbf{x}}[t_{\text{spinup}}, \dots, t_w]_{0.5\%} \quad (\text{A.2})$$

We use an additional period of time before the sequence phase to ensure that the sequence of states used to calculate the scaling is long enough to avoid problems of small-number statistics, while remaining short enough to avoid including variability on longer time scales than are represented in the sequence. In practice the pre-sequence was about 30% of the length of the sequence.

To generate the sequence of artificial observations \mathcal{X}_0 between $t = t_0$ and t_w we then add, at each t_i , a vector of random numbers $\tilde{\mathbf{e}}$ (noise) to the true state:

$$\mathbf{x}_{i,0} = \hat{\mathbf{x}}_i + \tilde{\mathbf{e}}_i \quad (\text{A.3})$$

The random numbers are independently and identically distributed (IID) and are originally drawn from the Gaussian distribution $N(0, 1)$. These random numbers are then converted to observational error by multiplying by a fixed fraction σ of the natural variability \mathbf{r} at that grid point,

$$\tilde{\mathbf{e}}_i \sim \sigma \mathbf{r} N(0, 1) \quad (\text{A.4})$$

The effect of the scaling by \mathbf{r} ensures that, statistically, the amount of noise added at each point and to each field is the same fraction (σ) of the natural variability.

Appendix A.4. Gradient descent applied to MORALS

To initialise the gradient descent algorithm, we first create a sequence of observations \mathcal{X}_0 of the true sequence $\hat{\mathcal{X}}$ using the method described in Sect. Appendix A.3 above.

The algorithm then enters its main loop, which advances the gradient descent by one iteration from step h to $h + 1$. The gradient descent loop begins with the sequence \mathcal{X}_h . The first step is to initialise MORALS using each state in the subsequence $\mathbf{x}_{i,h}$, $i \in \{0, \dots, w - 1\}$, and then to integrate MORALS forward by $\Delta t = t_{i+1} - t_i$ to obtain the forecast images $f(\mathbf{x}_{i,h})$, $i \in \{0, \dots, w - 1\}$. This is the *forecasting step*. The mismatch is then calculated for each state-forecast pair, except for the first state where there is no corresponding forecast. The mismatch is given by

$$\delta \mathbf{x}_{i,h} = \mathbf{x}_{i+1,h} - f(\mathbf{x}_{i,h}) \equiv \begin{pmatrix} \delta \mathbf{u}_{i,h} \\ \delta \mathbf{v}_{i,h} \\ \delta \mathbf{w}_{i,h} \\ \delta \mathbf{T}_{i,h} \end{pmatrix} \equiv \begin{pmatrix} \mathbf{u}_{i+1,h} \\ \mathbf{v}_{i+1,h} \\ \mathbf{w}_{i+1,h} \\ \mathbf{T}_{i+1,h} \end{pmatrix} - f \begin{pmatrix} \mathbf{u}_{i,h} \\ \mathbf{v}_{i,h} \\ \mathbf{w}_{i,h} \\ \mathbf{T}_{i,h} \end{pmatrix} \quad (\text{A.5})$$

for $i \in \{0, 1, \dots, w - 1\}$. The indeterminism of the sequence is then calculated, using a calculation based on Eq. (3). When applying this expression to MORALS, however, two problems must first be overcome.

First, it is not appropriate to add together the different MORALS fields because they represent different quantities expressed in different units. To combine the quantities in this way they should be expressed in a non-dimensional form. This is true even for a non-physical quantity like the indeterminism because otherwise I is poorly-defined - it depends on the units used for the different fields and its value can be changed just by changing those units even when the physical sequence of states itself has not changed. Second, the range of values in the four MORALS fields are quite different. \mathbf{T} is usually $O(1)$, \mathbf{u} and \mathbf{v} are $O(10^{-2})$, and \mathbf{w} is $O(10^{-3})$. The indeterminism combines the mismatch from all four fields, so the range of values in each field should be scaled before they are combined, otherwise the contribution to I from the velocity mismatches will be swamped by the contribution from the temperature mismatches.

Both these problems are solved by dividing the mismatches grid point-wise by the natural variability \mathbf{r} described in Appendix A.3. \mathbf{r} remains constant over the course of the gradient descent, so the scaling is the same for each iteration. This solves the first problem by converting each value into a dimensionless quantity, and it solves the second by dividing by a natural scale for each field and at each grid point. With this scaling, we define the mean squared indeterminism for a sequence \mathcal{X}_h to be

$$I(\mathcal{X}_h) = \frac{1}{wN} \sum_{i=0}^{w-1} \left\| \delta \mathbf{x}_{i,h} \circ \mathbf{r}^{-1} \right\|^2 \quad (\text{A.6})$$

where \circ denotes the Hadamard (pointwise) product. The indeterminism may also be calculated for a particular $\mathbf{x}_{i,h}$:

$$I(\mathbf{x}_{i,h}) = \frac{1}{N} \left\| \delta \mathbf{x}_{i,h} \circ \mathbf{r}^{-1} \right\|^2 \quad (\text{A.7})$$

and for a single field, say temperature:

$$I(\mathbf{T}_{i,h}) = \frac{1}{N_T} \left\| \delta \mathbf{T}_{i,h} \circ \mathbf{r}_T^{-1} \right\|^2 \quad (\text{A.8})$$

where \mathbf{r}_T denotes the temperature components of the \mathbf{r} vector. Because each field is scaled in the same way, comparing these quantities between fields provides information about which fields are contributing most to the overall mismatch. As I is a squared quantity, sums of these quantities also preserve the squared Euclidean norm:

$$I(\mathbf{x}_{i,h}) = \frac{N_u I(\mathbf{u}_{i,h}) + N_v I(\mathbf{v}_{i,h}) + N_w I(\mathbf{w}_{i,h}) + N_T I(\mathbf{T}_{i,h})}{N} \quad (\text{A.9})$$

$$I(\mathcal{X}_h) = \frac{1}{w} \sum_{i=0}^{w-1} I(\mathbf{x}_{i,h}) \quad (\text{A.10})$$

This definition of the indeterminism reflects its standard use by previous authors, as a mean over the model mismatches. The model grid itself is non-uniform; the model assigns higher grid resolution near the boundaries in order to resolve the boundary layers, but this is not reflected in the definition of the indeterminism as it should not be interpreted physically but as a purely mathematical construct.

Once $I(\mathcal{X}_h)$ has been calculated it is compared with a user-specified value ϵ . If $I(\mathcal{X}_h) \leq \epsilon$ then the gradient descent is terminated; ϵ is a parameter set by the user. If $I > \epsilon$ then a further check compares $I(\mathcal{X}_h)$ with $I(\mathcal{X}_{h-1})$.

Table A.3: Shadowing test parameters.

Best-case gradient descent	λ	0.25
Best-case iteration	h	9
Additional time	$t_{\text{end}} - t_w$	1500 s
Time of final state	t_{end}	3920 s
Number of Type I errors accepted	R	1
Number of candidate trajectories	E	130
Maximum number of significance tests in a trajectory	n	365

If $I(\mathcal{X}_h) > I(\mathcal{X}_{h-1})$ then the step length $\Delta\tau$ is halved, and the algorithm returns to the start of the *previous* iteration: $h \rightarrow h-1$. If $I(\mathcal{X}_h) \leq I(\mathcal{X}_{h-1})$ then $\Delta\tau$ is doubled for the next iteration. This adaptive refinement of the step length allows the algorithm to proceed quickly both in regions of the state space where I varies slowly with τ (adapting towards larger $\Delta\tau$) and where I varies rapidly (adapting towards smaller $\Delta\tau$). In our runs we disable doubling of the step length once a point is reached when $I(\mathcal{X}_h) > I(\mathcal{X}_{h-1})$. The effect of this feature is that in the first few gradient descent steps $\Delta\tau$ equilibrates to a value that causes I to fall smoothly in subsequent steps, while making optimum use of the available resources by running as few iterations as possible.

Finally \mathcal{X}_h is updated. Using the gradient-free descent method of *Judd et al. (2004)* with $\mathcal{A} = \lambda\mathbf{I}$, each new state is given by

$$\mathbf{x}_{i,h+1} = \mathbf{x}_{i,h} - \frac{2\Delta\tau}{w} \times \begin{cases} -\lambda\delta\mathbf{x}_{0,h} & i = 0 \\ \delta\mathbf{x}_{i-1,h} - \lambda\delta\mathbf{x}_{i,h} & 1 \leq i \leq w-1 \\ \delta\mathbf{x}_{w-1,h} & i = w \end{cases} \quad (\text{A.11})$$

where $\delta\mathbf{x}_{i,h} = \mathbf{x}_{i+1,h} - f(\mathbf{x}_{i,h})$ and λ is a scalar. \mathcal{X}_{h+1} is then the input for the next iteration of the gradient descent. The loop repeats while $I > \epsilon$ and $h \leq h_{\text{max}}$, where h_{max} is a maximum iteration number.

Appendix A.5. Significance level for the shadowing definition

We require a significance level p for the shadowing definition such that Type I errors are avoided for all the model trajectories. The largest significance level p such that in the event that the true candidate shadowing time is the trajectory length equivalent of n significance tests, fewer than R trajectories in a set of E candidates will suffer a Type I error, is (Part I, Eq. 14)

$$p = 1 - \left(1 - \frac{R}{E}\right)^{1/2n} \quad (\text{A.12})$$

We set n to the longest possible model trajectory in this context, $n = 365$. For the trajectory started from position $i = 0$, $n = 365$ comes from one significance test at lead time zero, $320/5 = 64$ tests over the sequence, and $1500/5 = 300$ over the extra period of observations. There are $E = 129$ candidates (not including candidates started from forecast images, which use the same data), and to ensure fewer than one Type I error throughout the whole sequence we require $R < 1$. Putting these into Eq. A.12 gives $p < 10^{-5}$.

Appendix A.6. Some comments on computational expense

The computational resources required to run the gradient descent algorithm are considerable, even without a full adjoint model. The procedure was partially parallelized by running on a multi-core computer, with the forecast stage split into blocks of four simulations at a time. Even then each iteration of the gradient descent took approximately 90 s (running on a single desktop computer with four Intel® Core™ 2 Q9400 CPUs running at 2.66GHz with 8GB RAM). 50–60% of this time was spent on the forecast stage and 30–40% setting up the parameter files for each simulation. Both of these steps require no cross-referencing from other simulations, so if run on a large cluster the computational overhead would be reduced significantly. The MORALS resolution is less than is normally used for simulations that are compared with laboratory data, which usually use $N_R = 24$, $N_\theta = 64$, $N_z = 24$ or higher (*Young and Read, 2008; Jacoby et al., 2011*). On a slightly older machine the $16 \times 32 \times 16$ run took 5 min per iteration, a test run at $24 \times 64 \times 24$ required about 15 min per iteration, and another test run at $8 \times 16 \times 8$ required only 90 s. The medium resolution is sufficient for demonstration purposes and because we are working in the PMS, but if laboratory data were used to initialise the gradient descent then a higher resolution would be required.

ARAKAWA, A., AND V. R. LAMB (1977), Computational Design of the Basic Dynamical Processes of the UCLA General Circulation Model, *Meth. Comput. Phys.*, 17:173–265, doi:10.1016/B978-0-12-460817-7.50009-4.

- BOWEN, R. (1975), omega-Limit Sets for Axiom A Diffeomorphisms, *J. Differ. Equations*, 18:333–339, doi:10.1016/0022-0396(75)90065-0.
- DU, H. (2009), Combining Statistical Methods with Dynamical Insight to Improve Nonlinear Estimation, Ph.D. thesis, London School of Economics.
- EVENSEN, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, 99:10,143–10,162, doi:10.1029/94JC00572.
- FARMER, J. D., AND J. J. SIDOROWICH (1991), Optimal shadowing and noise reduction, *Physica D*, 47:373–392, doi:10.1016/0167-2789(91)90037-A.
- FARNELL, L., AND R. PLUMB (1976), Numerical integration of flow in a rotating annulus II: three dimensional model, *Tech. rep.*, Occasional Note Met O 21 76/1, Geophysical Fluid Dynamics Laboratory, Meteorological Office, Bracknell, Berkshire.
- GILMOUR, I. (1998), ‘*t*-shadowing, probabilistic prediction and weather forecasting’, Ph.D. thesis, University of Oxford.
- GREBOGI, C., S. M. HAMMEL, J. A. YORKE, AND T. SAUER (1990), Shadowing of physical trajectories in chaotic dynamics: Containment and refinement, *Phys. Rev. Lett.*, 65:1527–1530, doi:10.1103/PhysRevLett.65.1527.
- HAMMEL, S. M. (1990), A noise reduction method for chaotic systems, *Phys. Lett. A*, 148:421–428, doi:10.1016/0375-9601(90)90493-8.
- HIDE, R. (1953), Some experiments on thermal convection in a rotating liquid, *Q. J. Roy. Meteor. Soc.*, 79:161, doi:10.1002/qj.49707933916.
- HIDE, R., AND P. MASON (1975), Sloping convection in a rotating fluid, *Adv. Phys.*, 24:47–100, doi:10.1080/00018737500101371.
- HIGNETT, P., A. A. WHITE, R. D. CARTER, W. D. N. JACKSON, AND R. M. SMALL (1985), A comparison of laboratory measurements and numerical simulations of baroclinic wave flows in a rotating cylindrical annulus, *Q. J. Roy. Meteor. Soc.*, 111:131–154, doi:10.1002/qj.49711146705.
- HUSSAIN, M. (2010), Tangent Linear and Adjoint Models for Fluid Flow in a Rotating Annulus, Master’s thesis, Johann Wolfgang Goethe University.
- IKEDA, K. (1979), Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, *Opt. Commun.*, 30:257–261, doi:10.1016/0030-4018(79)90090-7.
- JACOBY, T. N. L., P. L. READ, P. D. WILLIAMS, AND R. M. B. YOUNG (2011), Generation of inertia-gravity waves in the rotating thermal annulus by a localised boundary layer instability, *Geophys. Astro. Fluid*, 105:161–181, doi:10.1080/03091929.2011.560151.
- JUDD, K. (2003), Nonlinear state estimation, indistinguishable states, and the extended Kalman filter, *Physica D*, 183:273–281, doi:10.1016/S0167-2789(03)00180-5.
- JUDD, K. (2007), Failure of maximum likelihood methods for chaotic dynamical systems, *Phys. Rev. E*, 75:036,210, doi:10.1103/PhysRevE.75.036210.
- JUDD, K. (2008), Forecasting with imperfect models, dynamically constrained inverse problems, and gradient descent algorithms, *Physica D*, 237:216–232, doi:10.1016/j.physd.2007.08.017.
- JUDD, K., AND L. SMITH (2001), Indistinguishable states: I. Perfect model scenario, *Physica D*, 151:125–141, doi:10.1016/S0167-2789(01)00225-1.
- JUDD, K., AND L. A. SMITH (2004), Indistinguishable states II. The imperfect model scenario, *Physica D*, 196:224–242, doi:10.1016/j.physd.2004.03.020.
- JUDD, K., AND T. STEMLER (2009), Failures of sequential Bayesian filters and the successes of shadowing filters in tracking of nonlinear deterministic and stochastic systems, *Phys. Rev. E*, 79:066,206, doi:10.1103/PhysRevE.79.066206.
- JUDD, K., AND T. STEMLER (2010), Forecasting: it is not about statistics, it is about dynamics., *Philos. T. Roy. Soc. A*, 368:263–71, doi:10.1098/rsta.2009.0195.
- JUDD, K., L. SMITH, AND A. WEISHEIMER (2004), Gradient free descent: Shadowing, and state estimation using limited derivative information, *Physica D*, 190:153–166, doi:10.1016/j.physd.2003.10.011.
- JUDD, K., C. A. REYNOLDS, T. E. ROSMOND, AND L. A. SMITH (2008), The Geometry of Model Error, *J. Atmos. Sci.*, 65:1749–1772, doi:10.1175/2007JAS2327.1.
- KOSTELICH, E. J., AND J. A. YORKE (1988), Noise reduction in dynamical systems, *Phys. Rev. A*, 38:1649–1652, doi:10.1103/PhysRevA.38.1649.
- LORENC, A. C., R. S. BELL, AND B. MACPHERSON (1991), The Meteorological Office analysis correction data assimilation scheme, *Q. J. Roy. Meteor. Soc.*, 117:59–89, doi:10.1002/qj.49711749704.
- LORENZ, E. N. (1963), Deterministic nonperiodic flow, *J. Atmos. Sci.*, 20:130–141, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.
- RAVELA, S., J. MARSHALL, C. HILL, A. WONG, AND S. STRANSKY (2010), A realtime observatory for laboratory simulation of planetary flows, *Exp. Fluids*, 48:915–925, doi:10.1007/s00348-009-0752-0.
- RAWLINS, F., S. P. BALLARD, K. J. BOVIS, A. M. CLAYTON, D. LI, G. W. INVERARITY, A. C. LORENC, AND T. J. PAYNE (2007), The Met Office global four-dimensional variational data assimilation scheme, *Q. J. Roy. Meteor. Soc.*, 133:347–362, doi:10.1002/qj.32.
- READ, P. L., M. J. BELL, D. W. JOHNSON, AND R. M. SMALL (1992), Quasi-periodic and chaotic flow regimes in a thermally driven, rotating fluid annulus, *J. Fluid Mech.*, 238:599–632, doi:10.1017/S0022112092001836.
- READ, P. L., N. P. J. THOMAS, AND S. H. RISCH (2000), An evaluation of Eulerian and semi-Lagrangian advection schemes in simulations of rotating, stratified flows in the laboratory. Part I: Axisymmetric flow., *Mon. Weather Rev.*, 128:2835–2852, doi:10.1175/1520-0493(2000)128<2835:AEOEAS>2.0.CO;2.
- RIDOUT, D., AND K. JUDD (2002), Convergence properties of gradient descent noise reduction, *Physica D*, 165:26–47, doi:10.1016/S0167-2789(02)00376-7.
- SMITH, L. A. (2000), Disentangling uncertainty and error: On the predictability of nonlinear systems, in *Nonlinear Dynamics and Statistics*, edited by A. Mees, pp. 31–64, Birkhäuser Boston.
- SMITH, L. A., M. C. CUÉLLAR, H. DU, AND K. JUDD (2010), Exploiting Dynamical Coherence: A geometric approach to parameter estimation in nonlinear models, *Phys. Lett. A*, 374:2618–2623, doi:10.1016/j.physleta.2010.04.032.
- STEMLER, T., AND K. JUDD (2009), A guide to using shadowing filters for forecasting and state estimation, *Physica D*, 238:1260–1273, doi:10.1016/j.physd.2009.04.008.
- VAN LEEUWEN, P. J. (2010), Nonlinear Data Assimilation in geosciences: an extremely efficient particle filter, *Q. J. Roy. Meteor. Soc.*, 136:1991–1999, doi:10.1002/qj.699.
- YOUNG, R. M. B., AND P. L. READ (2008), Flow transitions resembling bifurcations of the logistic map in simulations of the baroclinic rotating annulus, *Physica D*, 237:2251–2262, doi:10.1016/j.physd.2008.02.014.
- YOUNG, R. M. B., AND P. L. READ (2013), Data assimilation in the laboratory using a rotating annulus experiment, *Q. J. Roy. Meteor. Soc.*, 139:1488–1504, doi:10.1002/qj.2061.
- YOUNG, R. M. B., AND P. L. READ (2016), Predictability of the thermally driven laboratory rotating annulus, *Q. J. Roy. Meteor. Soc.*, 142:911–927, doi:10.1002/qj.2694.
- YOUNG, R. M. B., R. BINTER, AND F. NIEHÖRSTER (2019), Shadowing the rotating annulus. Part I: Measuring candidate trajectory shadowing times, *arXiv*, physics.data-an:1909.04488.